Higher Horizons Tracking Report 2026 (October Update) Technical Annex:

statistical analysis of progression rates of Higher Horizons participants and non-participants

Dr. Hannah Merry

October 2026

www.higherhorizons.co.uk

Contents

Table of Figures 3
Introduction 4
Constructing the Analysis Dataset 5
Exclusion of Students with Missing Demographic Data 6
Exploratory Data Tables 7
Contact Hours Distribution 7
Progression Rates by Contact Hours Band 7
Activity Types Across Cohorts 8
Logistic Regression – Full Results and Diagnostics 8
Participation Model 9
Dose-Response Model 10
Activity Types Model 11
Propensity Score Methods – Full Results and Diagnostics 12
Participation Analysis 12
Dose-Response Propensity Score Analysis 13
Bayesian Analysis 14
Common specification 14
Participation Analysis 16
High Dose vs Low Dose 16
Dose-Response Curve 17
Activity Types 17

Table of Figures

Table 1: Summary of dataset filtering steps 5
Table 2: Missing data rates by variable and group 6
Table 3: Contact Hours Summary Statistics by Expected HE Entry Year 7
Table 4: Proportion of Participants in Each Contact Hour Band 7
Table 5: Progression rate by contact hours band and Expected HE Entry Year 7
Table 6: Proportion (%) of participants who attended at least one activity of each type, by Expected HE Entry Year 8
Table 7: Percentage of participants attending 1 / 2 / 3 / 4+ distinct activity types, by Expected HE Entry Year 8
Table 8: Logistic Regression Participation Model Results 9
Table 9: Year Stratified Participation Odds Ratio 10
Table 10: Logistic Regression Dose-Response Model Results 10
Table 11: Year-stratified dose-response odds ratios (vs No contact) 10
Table 12: Logistic Regression Activity Types Model Results 11
Table 13: Validation metrics for all logistic regression models 11
Table 14: Participation analysis – results from all methods 12
Table 15: Dose-response analysis – high dose (≥16h) vs low dose (<3h), participants only 14
Table 16: Bayesian model specifications 15
Table 17: Prior distributions 15
Table 18: Diagnostics for all Bayesian models 16

Introduction

This technical annex accompanies the Higher Horizons Tracking Report 2026 (October Update). It assumes familiarity with the methodology, dataset and findings described in that report, and provides the additional analytical detail that underpins those findings. The annex is intended for readers who wish to scrutinise or reproduce the analysis, such as researchers, evaluators, and technical staff.

This update adds the 2023 cohort to the analysis and makes a small number of changes to the methodology: disability is added as a demographic covariate; the early-progressor exclusion rule is made more targeted; and the "High" contact-hours band is redefined from 8 or more hours to 16 or more hours, with the dose-response propensity score analysis redefined to compare high-dose against low-dose participants. These changes mean the analysis is not a like-for-like comparison with the May 2026 report.

The annex covers five areas:

1. The full detail of dataset construction, including analysis of dropped students.
2. Exploratory data tables showing the distribution of contact hours, progression rates by contact hours band, and activity type engagement across cohort years.
3. Full logistic regression results for all models, including validation metrics and diagnostic checks.
4. Full propensity score matching results for both the participation and dose-response analyses including differences in how each analysis matched or dropped students.
5. Specification, priors and validation of the Bayesian models for participation, dose and activity types.

All tracking data used in this analysis was provided by HEAT. Thank you to the HEAT Service for their work in facilitating access to, and preparing, this dataset.

HESA data is Copyright Jisc 2026. Neither Jisc nor Jisc Services Limited can accept responsibility for any inferences or conclusions derived by third parties from data or other information supplied by Jisc or Jisc Services Limited. This analysis was conducted by Higher Horizons and its conclusions were derived by Higher Horizons.

The HESA Standard Rounding Methodology has been applied to all descriptive counts and percentages: counts are rounded to the nearest multiple of 5, and percentages or averages based on very small groups are suppressed. Statistical models were fitted on the underlying individual-level dataset without rounding.

Constructing the Analysis Dataset

The raw dataset contains 61,295 unique student records with expected HE entry years spanning 2015 to 2024. The filtering steps below were applied in sequence to produce the final analysis dataset. Table 1 summarises the cumulative effect of each step. Note: rounding is applied to rows independently so may not sum correctly.

Table 1: Summary of dataset filtering steps

Filtering step	Records removed	Remaining
Raw dataset	—	61,295
Restriction to 2019–2023 cohorts (excl. pre-2019 and 2024)	18,320	42,980
Exclusion of mature participants	330	42,645
Exclusion of mature non-participants	3,055	39,590
Removal of pre-expected entry progressors (fuzzy/partial matches only)	25	39,565
Exclusion of students missing sex, ethnicity, TUNDRA or IMD	2,765	36,800
Exclusion of students missing disability	3,120	33,680
Final analysis dataset	—	33,680

The cohort was first restricted to expected entry years 2019 to 2023, removing 18,320 students. Cohorts before 2019 contained very small numbers and few or no non-participants, so the effect of participation could not be separated from the baseline progression rate. The 2024 cohort was excluded because its by-19 tracking window is not yet complete; it requires HESA data for 2025/26, which is not yet available. Participants who first engaged with outreach after their expected HE entry year were then removed (330 students), as they had passed the outcome window before taking part. Non-participants whose records were collected after their expected entry year were also removed (3,055 students). These students were still at a Higher Horizons target school or college after they would have been expected to progress, and including them would have biased the non-participant progression rate downwards.

The data collection date was not used to filter participants. In HEAT this is an administrative field that records when data was received or last updated, not when the student took part. Participants who would have been removed by this filter progressed at a much higher rate than the equivalent non-participants (35.3% against 8.6%), which indicates that their collection dates are unreliable. Activity dates were used instead, since they already exclude participants who genuinely engaged after the outcome window. Finally, 25 students recorded as entering HE before their expected entry year were removed. Early entry is possible in a small number of cases, so students matched exactly to the HESA record on name, date of birth and postcode were retained. Only those relying on a fuzzy or partial match were removed, as a weaker match is more

likely to reflect a data error. The previous report removed all such records regardless of match quality. Students with missing demographic data were then excluded, as described below.

Exclusion of Students with Missing Demographic Data

The statistical methods all require complete data for every covariate. This year five variables are used as covariates: sex, ethnicity (grouped), disability, TUNDRA quintile, and IMD quintile. Table 2 shows the rate of missing data for each field.

Table 2: Missing data rates by variable and group

Variable	Participants	Non-participants
Tundra MSOA Quintile	0.42%	2.04%
Ethnic Group	2.57%	6.01%
Sex	2.80%	1.99%
IMD Quintile	0.04%	0.19%
Free School Meals	12.90%	33.04%
Disability	9.27%	9.58%

Free school meal eligibility and disability status both have far higher rates of missing data than the other fields. Free school meal eligibility was therefore not used as a covariate. Disability, despite around 9% of records missing this field, was retained as a covariate this year because it is a meaningful predictor of progression; this is a change from the previous report, in which disability was not used. Students missing sex, ethnicity, TUNDRA or IMD were excluded first (removing around 2,765 students), and students missing disability were then excluded (removing a further 3,120 students), giving a final analytical sample of 33,680.

The students dropped for missing demographic data progress to higher education at roughly the same rates as the retained sample. The dropped participant progression rate is 2.4pp lower than the retained sample, while the dropped non-participant progression rate is 0.3pp higher than the retained sample. This slight unevenness in the differences between the progression rates means that the progression gap between the dropped participants and dropped non-participants is slightly smaller (+5pp) than that between the retained participants and retained non-participants (+7.7pp).

Exploratory Data Tables

This section provides detailed data tables supporting the exploratory analysis presented in the summary report.

Contact Hours Distribution

Table 3: Contact Hours Summary Statistics by Expected HE Entry Year

	2019	2020	2021	2022	2023
Mean	6.8	6.8	11.4	13.6	12.7
Median	6.0	6.0	7.0	8.0	6.5
Minimum	0.5	0.4	0.3	0.2	0.5
Maximum	63.0	66.0	134.0	135.5	162.0

Participants were grouped into three contact hours bands using the revised banding: Low (fewer than 3 hours), Medium (at least 3 and fewer than 16 hours), and High (16 or more hours). Table 4 shows the proportion of participants in each intensity band by expected HE entry year.

Table 4: Proportion of Participants in Each Contact Hour Band

Expected Entry Year	Low	Medium	High
2019	30.8%	60.1%	9.0%
2020	28.1%	64.6%	7.4%
2021	23.5%	52.9%	23.6%
2022	18.7%	50.3%	31.0%
2023	22.9%	50.6%	26.5%
Total	23.6%	54.0%	22.4%

Progression Rates by Contact Hours Band

Table 5 shows raw progression rates by band and cohort year. The increase in progression rate from Low to Medium to High holds within every cohort year.

Table 5: Progression rate by contact hours band and Expected HE Entry Year

Year	No contact	Low (<3h)	Medium (≥3, <16h)	High (≥16h)
2019	41.6%	42.9%	55.5%	59.6%
2020	36.7%	38.3%	45.4%	50.6%
2021	27.1%	34.8%	38.5%	39.0%
2022	28.8%	31.5%	34.0%	37.2%
2023	30.8%	30.7%	32.9%	37.4%

Activity Types Across Cohorts

Table 6 shows the proportion of participants in each cohort who attended at least one activity of each type. IAG was the most common activity type throughout, growing from 53% of participants in 2019 to 82% in 2023. Campus Visit reach was broadly stable across all years.

Table 6: Proportion (%) of participants who attended at least one activity of each type, by Expected HE Entry Year

Year	Campus Visit	IAG	Attainment Raising	Subject Specific	Mentoring	Summer School
2019	40.6%	53.0%	10.5%	8.4%	2.6%	3.6%
2020	44.0%	55.3%	25.7%	15.4%	2.9%	0.8%
2021	39.6%	76.9%	26.6%	16.4%	2.7%	7.0%
2022	44.5%	76.6%	29.2%	28.6%	3.4%	4.3%
2023	40.7%	81.9%	19.6%	28.0%	1.0%	1.1%

Table 7: Percentage of participants attending 1 / 2 / 3 / 4+ distinct activity types, by Expected HE Entry Year

Year	1	2	3	4+
2019	83.5%	13.7%	2.2%	0.6%
2020	64.1%	27.9%	7.5%	0.5%
2021	52.8%	29.9%	12.7%	4.6%
2022	47.9%	27.5%	15.2%	9.3%
2023	54.0%	23.9%	16.6%	5.5%
Grand Total	57.1%	25.4%	12.5%	5.0%

The proportion of students only experiencing one activity type across the period fell from 84% in 2019 to 48% in 2022 before rising slightly to 54% in 2023. Experiencing 3 or more activity types increased across the period, rising from just 3% in 2019 to 22% in 2023.

Logistic Regression – Full Results and Diagnostics

Three logistic regression models were fitted in sequence. This section presents the full results tables, validation metrics and diagnostic checks for each. All models were fitted on the same 33,680 individual-level records.

Four standard diagnostics are reported for each model plus a maximum calibration deviation. The AUC-ROC measures discrimination, with 0.5 indicating chance and values of 0.7 or above considered acceptable. The Brier Score measures average

prediction error, with lower scores indicating a better fit. The Hosmer–Lemeshow test assesses calibration; a p-value above 0.05 indicates no evidence of poor calibration. The Pseudo R-squared is a relative measure of fit. As with any behavioural outcome, AUC values below conventional thresholds are expected; calibration is the more relevant diagnostic for estimating a participation effect.

Participation Model

This model uses expected entry year, sex, ethnicity (grouped), disability, TUNDRA quintile and IMD quintile as covariates. This is the primary model for estimating the adjusted participation effect.

Table 8: Logistic Regression Participation Model Results

Term	Odds Ratio	95% Confidence Interval	p-value
Participated (vs not)	1.440	1.368 – 1.515	<0.001
Year 2020 (vs 2019)	0.801	0.721 – 0.890	<0.001
Year 2021 (vs 2019)	0.593	0.539 – 0.652	<0.001
Year 2022 (vs 2019)	0.578	0.527 – 0.633	<0.001
Year 2023 (vs 2019)	0.596	0.545 – 0.651	<0.001
Sex (Male vs Female)	0.539	0.514 – 0.565	<0.001
Disability (Yes vs No)	0.602	0.566 – 0.640	<0.001
Ethnicity – Asian (vs White)	2.747	2.479 – 3.044	<0.001
Ethnicity – Black (vs White)	1.679	1.412 – 1.996	<0.001
Ethnicity – Mixed (vs White)	1.373	1.219 – 1.545	<0.001
Ethnicity – Arab (vs White)	0.659	0.393 – 1.104	0.113
Ethnicity – Other (vs White)	0.699	0.479 – 1.021	0.064
TUNDRA Q2 (vs. 1)	0.978	0.915 – 1.046	0.522
TUNDRA Q3 (vs. 1)	1.143	1.062 – 1.229	<0.001
TUNDRA Q4 (vs. 1)	1.287	1.184 –1.400	<0.001
TUNDRA Q5 (vs. 1)	1.410	1.273 –1.562	<0.001
IMD Q2 (vs. 1)	1.209	1.124 –1.300	<0.001
IMD Q3 (vs. 1)	1.575	1.457 –1.703	<0.001
IMD Q4 (vs. 1)	1.742	1.606 –1.889	<0.001
IMD Q5 (vs. 1)	1.951	1.788 –2.129	<0.001

A multicollinearity check found the highest variance inflation factors among the cohort-year indicators (all below 4). TUNDRA and IMD quintile are correlated (Pearson $r \approx 0.61$) but re-running the model with each in turn confirmed the participation odds ratio is essentially unchanged (1.44 with both, 1.39 with TUNDRA only, 1.39 with IMD only), so the collinearity does not distort the headline estimate. TUNDRA and IMD were entered as five-level categorical variables, so no linearity assumption is required.

Expressed as an ATT via g-computation, the model's participation estimate is **+7.6 percentage points** (95% CI from bootstrapping 500 iterations is +6.7 to +8.7), which places the regression result within the range of the propensity score estimates below. The E-value for the participation effect is 1.85 (1.76 at the confidence-interval limit).

The participation OR was positive and statistically significant in each cohort year.

Table 9: Year Stratified Participation Odds Ratio

Cohort	Participation OR	95% CI	p-value
2019	1.63	1.36–1.96	< 0.001
2020	1.45	1.22–1.72	< 0.001
2021	1.80	1.59–2.05	< 0.001
2022	1.43	1.30–1.58	< 0.001
2023	1.27	1.17–1.38	< 0.001

A sensitivity check adding free school meal eligibility as a three-category variable (Yes / No / Unknown) increased the participation odds ratio to 1.59 and the g-computation ATT to +9.5pp (8.6pp – 10.6pp). This did not change the direction or significance of the finding. Because free school meal eligibility could not be included as a clean binary without dropping around a quarter of the sample, and because its "Unknown" category is itself associated with higher progression, it was not added to the primary model.

Dose-Response Model

The dose-response model replaces the binary participation indicator with a four-category contact hours measure: No contact (reference), Low (<3 hours), Medium (≥3 and <16 hours), High (≥16 hours).

Table 10: Logistic Regression Dose-Response Model Results

Contact hours band	Odds Ratio	95% Confidence Interval	p-value
Low (<3h) vs No contact	1.127	1.043 – 1.217	0.002
Medium (≥3 and <16h) vs No contact	1.472	1.388 – 1.562	<0.001
High (≥16h) vs No contact	1.815	1.676 – 1.964	<0.001

Table 11 shows the year-stratified dose-response odds ratios. The Low < Medium < High ordering holds in every cohort year. The Medium and High bands are statistically significant in all five years; the Low band is significant in only two of the five (2021 and 2022). This suggests the effect of a low dose is not reliably detectable in most years.

Table 11: Year-stratified dose-response odds ratios (vs No contact)

Band	2019	2020	2021	2022	2023
Low	1.13	1.14	1.45	1.20	1.02
Medium	1.85	1.57	1.87	1.38	1.26
High	2.64	2.01	2.22	1.76	1.61

Activity Types Model

This model replaces the single participation indicator with separate "ever attended" flags for each of the seven activity types. Each odds ratio is the association with progression for attending at least one activity of that type, holding cohort year, demographics and attendance at other activity types constant.

Table 12: Logistic Regression Activity Types Model Results

Activity type	N attended	% of sample	Odds Ratio	95% CI	p-value
Campus Visit	7,390	21.9%	1.541	1.452 – 1.634	<0.001
Other	145	0.4%	1.714	1.209 – 2.429	0.003
Summer School	595	1.8%	1.377	1.158 – 1.638	<0.001
IAG	12,825	38.1%	1.199	1.139 – 1.262	<0.001
Mentoring Programme	425	1.3%	1.259	1.025 – 1.545	0.028
Subject Specific	3,845	11.4%	0.976	0.902 – 1.057	0.549
Attainment Raising	4,110	12.2%	0.940	0.871 – 1.016	0.118

The Attainment Raising result is a notable change from the May 2026 report, in which it was significantly negative (OR 0.85). This year it is not statistically significant (OR 0.94, $p = 0.118$). A year-stratified version of the model shows the Attainment Raising odds ratio rising across cohorts, from around 0.67 in 2019 to around 1.12 in 2023, and a formal likelihood-ratio test confirmed that the activity-type effects vary significantly by year ($p < 0.001$). This indicates the pooled change reflects the changing make-up of later cohorts. Because prior attainment is not available in this dataset, and Attainment Raising activities are typically targeted at lower-attaining students, this result cannot be fully interpreted and is presented as a secondary, exploratory finding only.

Table 13: Validation metrics for all logistic regression models

Metric	Participation Model	Dose Response Model	Activity Types Model
AUC-ROC	0.660	0.662	0.663
Brier Score	0.210	0.209	0.209
Hosmer–Lemeshow p-value	0.871	0.568	0.652
Pseudo R-squared	0.057	0.059	0.060
Max calibration deviation (percentage points)	0.7	1.0	1.1

Propensity Score Methods – Full Results and Diagnostics

This section provides the full analytical detail for the propensity score analyses summarised in the summary report. The propensity score for each student was estimated using a logistic regression predicting participation from sex, ethnicity, disability, TUNDRA quintile and IMD quintile, fitted separately within each cohort year. The matching specification (a caliper of 0.2 standard deviations of the logit of the propensity score) is unchanged from the previous report.

Participation Analysis

Four methods were applied: propensity score matching without replacement (the reportable matched estimate), coarsened exact matching, inverse probability of treatment weighting, and a propensity score stratification check. Table 14 summarises the results.

Table 14: Participation analysis – results from all methods

Method	ATT	95% CI	Analysed	Drop rate
Logistic regression (participation model, g-computation)	+7.6pp	+6.7 to +8.7	33,680	0%
Propensity score matching	+7.2pp	+6.0 to +8.4	11,170 pairs	36.8%
Coarsened exact matching	+8.0pp	+6.7 to +9.3	17,125 treated	3.1%
Inverse probability weighting	+7.5pp	+6.3 to +8.7	33,680	0%
Propensity score stratification	+7.9pp	+6.7 to +9.1	33,680	0%

Propensity Score Matching (PSM). 1:1 nearest-neighbour matching without replacement, exact on cohort year. This produced 11,170 matched pairs from 17,670 participants, an overall match rate of 63.2%. Match rates were lowest in the early cohorts (32.3% in 2019, 37.2% in 2020) and highest in 2023 (88.3%), reflecting the growing non-participant control pool. Post-matching balance passed on all covariates (standardised mean differences below 0.1, variance ratios within 0.8–1.25). The ATT was +7.2 percentage points (95% CI +6.0 to +8.4, $p < 0.001$ using McNemar's test), positive in every cohort year. A Monte Carlo check across 200 random matching orders confirmed the estimate is highly stable (standard deviation 0.2 percentage points).

The 36.8% drop rate exceeds the conventional 10% threshold, so the PSM ATT applies primarily to the matched subset rather than all participants. As in the previous report, the dropped participants were systematically more disadvantaged than those retained (mean TUNDRA quintile 1.66 versus 2.32, mean IMD quintile 1.97 versus 2.76, and more

concentrated in ethnic minority groups). This is the reason coarsened exact matching and inverse probability weighting, which retain far more participants, are given weight alongside PSM. The drop rate is lower than the 43% reported last year, reflecting the larger 2023 control pool rather than any change in matching specification.

A Rosenbaum bounds analysis gave a critical Γ of 1.35: an unobserved confounder would need to make a student have 1.35 times the odds of participating, holding observed demographics constant, to reduce the ATT to non-significance. An E-value of 1.79 (1.68 at the confidence-interval limit) was calculated from a matched risk ratio of 1.24. Both indicate a moderate level of robustness, consistent with interpreting the results as associations rather than confirmed causal effects.

Coarsened Exact Matching (CEM). CEM matched participants to non-participants sharing exactly the same profile across cohort year, sex, ethnicity, disability, TUNDRA quintile and IMD quintile, using variable match ratios so that no participant is dropped unless no exact match exists. It matched 17,125 of 17,670 participants (96.9%), dropping only 3.1%. Balance is exact by construction. The ATT was +8.0 percentage points (95% CI +6.7 to +9.3, $p < 0.001$), positive in every cohort year.

Inverse Probability of Treatment Weighting (IPTW). IPTW reweights the non-participant group so its weighted demographic distribution matches the participant group, excluding no one. Weight diagnostics were well behaved, and the ATT was stable across raw, trimmed and stabilised weight variants. The ATT was +7.5 percentage points (95% CI +6.3 to +8.7, $p < 0.001$), positive in every cohort year.

Propensity Score Stratification. As a further check, students were sorted into five propensity quintiles within each cohort year and the effect computed within each. The stratified ATT was +7.9 percentage points (95% CI +6.7 to +9.1), positive in all 25 year-by-quintile cells. Within-quintile balance was imperfect, mainly on TUNDRA quintile (maximum |SMD| 0.13–0.21), which is why stratification is treated as a corroborating method rather than the headline.

Dose-Response Propensity Score Analysis

The dose-response analysis has been redefined this year. Rather than comparing substantive engagement (3 or more hours) against little or no contact, it compares high-dose participants (16 or more contact hours) directly against low-dose participants (fewer than 3 hours), using participants only. Medium-dose participants are excluded so the contrast is between the two extremes of the dose distribution. This answers the more tightly specified question of whether, among participants, more contact makes more difference. The treatment group comprises 3,960 high-dose participants and the control group 4,170 low-dose participants.

Because both the treatment threshold and the comparison group differ from the previous report, these figures are *not* directly comparable to last year's dose estimate. Dose is also not randomly assigned — students accumulate more hours partly through

their school's engagement, which may also predict progression. Table 15 summarises the results.

Table 15: Dose-response analysis – high dose (≥16h) vs low dose (<3h), participants only

Method	ATT	95% CI	Analysed	Drop rate
Propensity score matching	+10.2pp	+7.5 to +12.9	2,460 pairs	37.8%
Coarsened exact matching	+10.6pp	+8.0 to +13.3	3,840 treated	3.1%
Inverse probability weighting	+10.7pp	+8.2 to +13.2	8,130	0%
Propensity score stratification	+10.4pp	+7.9 to +12.8	8,130	0%

Propensity score matching produced 2,460 matched pairs (a 62% match rate among high-dose participants) with an ATT of +10.2 percentage points (95% CI +7.5 to +12.9, $p < 0.001$), positive in every cohort year. Coarsened exact matching retained 96.9% of high-dose participants and gave +10.6 percentage points (95% CI +8.0 to +13.3). Inverse probability weighting included all participants in the two dose groups and gave +10.7 percentage points (95% CI +8.2 to +13.2). The propensity score stratification check gave +10.4 percentage points, though within-quintile balance failed on the ordinal quintile variables, so it is treated as corroboration only.

The three main methods produce estimates within 0.5 percentage points of each other (+10.2 to +10.7 pp). Among participants, those receiving 16 or more contact hours progressed around 10 percentage points more often than those receiving fewer than 3 hours, after accounting for demographics and cohort. As with all the propensity score analyses, prior attainment remains the most plausible uncontrolled variable, and these findings should be interpreted as associations rather than confirmed causal effects.

Bayesian Analysis

Four Bayesian logistic regression models were fitted. This section sets out how they were specified and validated. Estimates are posterior means with 95% equal-tailed credible intervals (CrI).

Common specification

All four models use the same covariates as the logistic regression models, except that TUNDRA and IMD quintiles are entered as linear terms. Entering either as categorical changed the participation ATT by at most 0.1pp. Each model has a binomial likelihood with a logit link, fitted to counts in each unique covariate combination (Table 16). This gives the same posterior as a model fitted to individual records. Each was sampled with NUTS: four chains of 2,000 draws after 2,000 warm-up iterations.

Effects are converted to percentage points by g-computation. For each posterior draw, the probability of progression is predicted for each student in the final column of Table

16 with the exposure switched on and off, and the difference is averaged. For the dose-response curve, predictions are made at each level of contact hours instead. The credible interval therefore carries the uncertainty in every coefficient.

Table 16: Bayesian model specifications

Model	Sample	Cells	Exposure	Effect averaged over
Participation	All 33,680 students	2,148	Participated vs not	17,670 participants
High vs low dose	8,130 participants (≥16h or <3h)	1,343	≥16h vs <3h	3,960 high-dose participants
Dose-response curve	17,670 participants	8,540	Contact hours (log scale)	17,670 participants
Activity types	17,670 participants	4,976	Seven yes/no type indicators	17,670 participants

The participation and dose effects have vague priors, so those estimates come from the Higher Horizons data alone. The sex and ethnicity priors are centred on Department for Education progression-by-19 rates for the six Higher Horizons local authorities over the same entry years, weighted as if worth 300 students. The benchmark TUNDRA and IMD gradients were two to three times steeper than in this sample, so those terms have vague priors (Table 17).

Table 17: Prior distributions

Term	Prior	Basis
Participation or dose effect	Normal(0, 1.5)	Vague
Intercept	Normal(logit of observed rate, 1.5)	Each model's observed progression rate
Sex (male vs female)	Normal(−0.56, 0.17)	Benchmark: female 49.3%, male 35.7%
Ethnicity – Asian / Black / Mixed (vs White)	Normal(0.85 / 0.70 / 0.26, 0.17)	Benchmark: 62.1% / 58.5% / 47.6%, White 41.2%
All other covariates	Normal(0, 1)	Vague

Every fit had to pass a convergence gate before results were extracted: maximum R-hat of 1.01 or below, minimum bulk and tail effective sample size (ESS) of at least 400, and no divergences. Models were compared by leave-one-out cross-validation (LOO). Because the data are aggregated, this leaves out one covariate cell at a time, which is valid for comparing models fitted to the same cells. Overlap is the share of students

whose covariate cell contains no comparison student, where the estimate relies on extrapolation (Table 18).

Table 18: Diagnostics for all Bayesian models

Model	Max R-hat	Min bulk / tail ESS	Max Pareto-k	Weight in cells with no comparison student
Participation	1.00	2,952 / 3,767	0.31	3.1%
High vs low dose	1.00	2,637 / 3,638	0.30	3.1% (a further 13.5% with 1–4)
Dose-response curve	1.00	2,128 / 2,743	0.23	Low 3.1%, Medium 1.9%, High 8.7%
Activity types	1.00	1,033 / 1,747	0.19	–

Participation Analysis

The participation odds ratio was 1.430 (95% CrI 1.360 to 1.504) and the ATT +7.5pp (+6.5 to +8.5). The other odds ratios were close to the regression estimates in Table 8. The posterior predictive progression rate was 0.345 (0.338 to 0.353) against an observed 0.346, and observed and predicted rates agreed within 1.6pp at every level of every covariate.

Varying the benchmark weight from 100 to 5,000 students moved the ATT by at most 0.06pp, and vague sex and ethnicity priors gave +7.48pp against +7.49pp. Leaving out each cohort gave +6.6pp to +9.1pp. A model with a participation × cohort year interaction predicted better in LOO (probability the primary model is worse: 0.94) and gives the year estimates in Table 4 of the summary report.

The E-value was calculated for each posterior draw, converting the odds ratio to a risk ratio at participants' estimated progression rate had they not taken part. It was 1.84 (95% CrI 1.74 to 1.94). The tipping-point analysis applied the standard bias adjustment for a binary confounder to each draw of the participation log-odds ratio. The credible interval first included zero when the confounder doubled the odds of progression with an 80/20 participant/non-participant split, or multiplied them by 2.5 with a 70/30 split.

High Dose vs Low Dose

The high-dose odds ratio was 1.68 (95% CrI 1.51 to 1.86) and the ATT +10.7pp (+8.6 to +12.9). Restricting the ATT to the 3,840 high-dose students in cells with a low-dose student gave the same +10.7pp (+8.6 to +12.8), so the weaker overlap does not drive the result. Vague sex and ethnicity priors gave +10.72pp against +10.74pp. The E-value was 2.16 (1.94 to 2.38). The credible interval first included zero at a confounder odds ratio of 2.5 with an 80/20 split, or 4 with a 70/30 split. The higher E-value reflects the larger effect, not greater robustness.

Dose-Response Curve

The reported model is a linear spline in log hours, with knots fixed at 3 and 16 hours. The base slope has a Normal(0, 1) prior and the two changes in slope have Normal(0, 0.5) priors, so the model assumes no change unless the data show one. It was compared with a banded model and a model with one slope in log hours. In LOO, the spline predicted best (stacking weight 0.95) and the single-slope model was clearly worse (probability worse 0.99), so no per-doubling figure is reported.

The slope in log-odds per log-hour was +0.28 (+0.18 to +0.38) below 3 hours, +0.01 (−0.06 to +0.08) from 3 to 16 hours, and +0.49 (+0.37 to +0.61) above 16 hours. Both changes in slope had credible intervals excluding zero. Because the knots were fixed, the flat middle section is exploratory. Support is thinnest at high hours, so the curve is reported only up to the 95th percentile of 37 hours. The banded model gave the standardised band rates in the summary report.

Activity Types

The reported model uses partial pooling. The seven type effects share a Normal(0, τ) distribution, with τ ~ HalfNormal(0.5), which pulls types attended by few students towards zero. The distribution is centred on zero because every participant attended at least one type, so an average type effect cannot be separated from the intercept. A version estimating that average failed the convergence gate, and centring on zero changed the type estimates by less than 0.01 log-odds. A comparison model with independent Normal(0, 1) priors predicted equally well (ELPD difference 0.0, SE 1.0) and also passed the gate. The pooled model is reported because it limits the influence of the smallest types, not because it predicts better.

τ was 0.29 log-odds (0.15 to 0.56). Pooling reduced the estimate for the smallest type, Other, by 0.147 log-odds and for IAG by only 0.004. Risk differences, P(>0) and P(best) are in Table 7 of the summary report.